Multiple Linear Regression with Qualitative
Independent Variables

1. Regression Equation

This remains the same as before. For example, suppose we have the following data with two predictors, student sex and teacher:

Table 1

Math Scores	Student Sex	Teacher	Math Scores	Student Sex	Teacher	Math Scores	Student Sex	Teacher
72	F	Gunther	74	F	Bryan	78	F	Marijke
73	F	Gunther	75	F	Bryan	79	F	Marijke
74	F	Gunther	76	F	Bryan	80	F	Marijke
76	M	Gunther	80	M	Bryan	83	M	Marijke
77	M	Gunther	81	M	Bryan	84	M	Marijke
78	M	Gunther	82	M	Bryan	85	M	Marijke

These data are plotted here:

http://tinyurl.com/29vcsep

or

https://spreadsheets.google.com/ccc?key=0ArHM99WFArnmdFhUMVNqVkNzN1ItX0JHWGxoRVFoU3c&hl=en &authkey=CJK3vZoJ

These data may be downloaded here:

http://www.bwgriffin.com/gsu/courses/edur8132/notes/math_scores.sav

The sample regression equation may take this form:

$$Y_i = b_0 + b_1 Male_{1i} + b_2 Bryan_{2i} + b_3 Marijke_{3i} + e_i, \qquad (1)$$

Regression coefficients maintain interpretations as learned previously:

b_1 = since Male will be dummy variable, b_1 is mean difference in math scores between males and females controlling for teacher.

b_2 = dummy variable for teacher Bryan, b_2 is the mean difference in math scores between Bryan and Gunther (the omitted or reference teacher) controlling for student sex.

b_3 = dummy variable for teacher Marijke, b_3 is the mean difference in math scores between Marijke and Gunther controlling for sex.

b_0 = predicted value of Y, Y', when IV equal zero; note that when dummy variables are in the equation, values of 0 for dummy represent the omitted group; literal interpretation for b_0 in this equation:

b_0 is the predicted mean math score for females in Gunther's class.

Additional Example of Interpretation of b0 when categorical IV present:

http://tinyurl.com/2wl9ssy or
https://spreadsheets.google.com/ccc?key=0ArHM99WFArnmdE51Z1VNWFhIdXZ3Qld3NlJldm0xVWc&hl=en

2. Predicted Values and Errors

As before, predicted values are obtained using the equation:

$Y' = b_0 + b_1Male_{1i} + b_2Bryan_{2i} + b_3Marijke_{3i}$ (2)

Residuals are obtained by

$e_i = Y - Y'$.

For the current data the following results are obtained:

Y' = 72.5 + 5.00 (Male) + 3.00 (Bryan) + 6.50 (Marijke)

(1) What is the predicted mean score for Females in Gunther's class?
(2) What is the predicted mean score for Males in Gunther's class?
(3) What is the predicted mean score for Females in Bryan's class?
(4) What is the predicted mean score for Males in Bryan's class?
(5) What is the predicted mean score for Females in Marijke's class?
(6) What is the predicted mean score for Males in Marijke's class?

(7a) What is the estimated student sex difference in math holding constant teacher?
(7b) Is this difference the same for all teachers? How does the average/estimated difference compare with actual?
(8) What are the estimated teacher differences in math holding constant student sex? How do these differences compare across student sex (compare estimated vs observed differenes)?

3. Predicted Values Holding Constant One IV

If one wished to obtain the predicted means for each teacher controlling for sex – not predicting means separately for males and females, but instead holding constant sex—one must include sex in the regression equation but instead of using the scors 0, 1, one instead using the mean value of sex.

In this example sex M = 0.50, so:

$Y' = b_0 + b_1(0.5) + b_2Bryan_{2i} + b_3Marijke_{3i}$ (2)

(7) What is the predicted mean score for Gunther's class?
(8) What is the predicted mean score for Bryan's class?
(9) What is the predicted mean score for Marijke's class?

4. Overall Mode Fit and Statistical Inference

The usual statistics apply for overall model fit (R^2, adjusted R^2, MSE, SEE, F-value).

5. Individual IV Statistical Inference

As before, each regression coefficient is tested with a t-ratio (b/se = t).

However, to test the *Global Effect* (overall statistical effect on the regression model) of a categorical IV with more than two categories, one must calculate $\Delta R^2(X_k)$ for that variable then perform the normal partial F test on $\Delta R^2(X_k)$.

Current Example: Teacher Global Effect:

Table 2

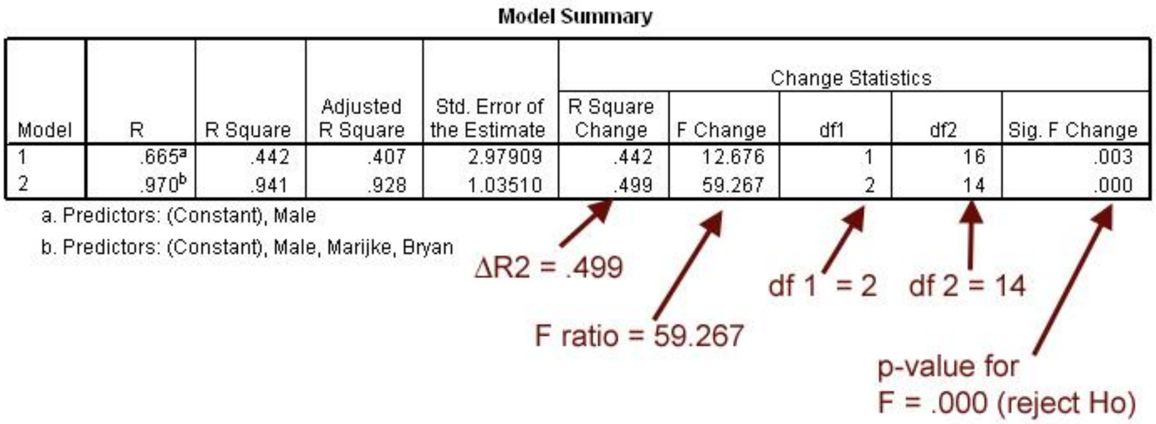
Model	R^2	Regression df	Error df
$Y' = b_0 + b_1Male_{1i}$	.442	1	16
$Y' = b_0 + b_1Male_{1i} + b_2Bryan_{2i} + b_3Marijke_{3i}$	.941	3	14
ΔR^2(Teacher) =	.941-.442 = .499	ΔR^2 df_1 = 3-1 = 2	ΔR^2 df_2 = 14 (smaller df)

In SPSS:
1. Choose Regression, enter Math in the Dependent box
2. Enter Male in Independents box, then click on Statistics->R-square Change->Continue

3. Click Next, then enter Bryan and Marijke dummy variables in IV box
4. Click Ok

See image below for SPSS results showing test of global effect ΔR^2(Teacher).

Figure 1



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.665[a]	.442	.407	2.97909	.442	12.676	1	16	.003
2	.970[b]	.941	.928	1.03510	.499	59.267	2	14	.000

a. Predictors: (Constant), Male
b. Predictors: (Constant), Male, Marijke, Bryan

6. Pairwise Comparisons Among IV Categories

It is possible to control familywise Type 1 error rate and perform pairwise comparisons. One may be interested in learning whether differences exist across teachers holding constant student sex differences in math scores. To perform these comparison follow these steps:

1. Estimate teacher differences in regression. To obtain all mean differences you will have to change reference groups in the regression equation, for example:

$Y' = b_0 + b_1 Male_{1i} + b_2 Bryan_{2i} + b_3 Marijke_{3i}$

b_2 = Bryan vs. Gunther mean difference
b_3 = Marijke vs. Gunther mean difference

How to get Bryan vs. Marijke mean difference? Rerun regression with Marijke as the omitted, referenced teacher:

$Y' = b_0 + b_1 Male_{1i} + b_2 Bryan_{2i} + b_3 Gunther_{3i}$

Now,
b_2 = Bryan vs. Marijke mean difference

2. Find standard errors (se) for each mean difference
3. Find appropriate Bonferroni or Scheffé critical t-value
4. Calculate CI for each mean difference, e.g,

Upper CI: b2 + se * critical t
Lower CI: b2 – se * critical t

For the current example Bonferroni critical t = 2.709 (with comparisons = 3 and df = 14). For the Bryan vs. Gunther comparison (b2 = 3.00, se = 0.598), the 95% Bonferroni CI is

Upper CI: 3.00 + 0.598 * 2.709 = 4.62
Lower CI: 3.00 – 0.598 * 2.709 = 1.38

Below is a table showing complete results.

Table 3

Comparison	Estimated Mean Difference	Standard Error of Difference	Bonferroni Adjusted 95% CI
Bryan vs. Gunther	3.00*	0.598	1.38, 4.62
Marijke vs. Gunther	6.50*	0.598	4.88, 8.12
Bryan vs. Marijke	-3.50*	0.598	-5.12, -1.88

*p<.05, where p-values are adjusted using the Bonferroni method.

7. APA Style Results

Table 4
Descriptive Statistics and Correlations Among Math Scores, Student Sex, and Teachers

Variable	Correlations			
	Math Scores	Male	Bryan	Marijke
Math Scores	---			
Male	.67*	---		
Bryan	-.03	.00	---	
Marijke	.63*	.00	-.50*	---
Mean	78.17	0.50	0.33	0.33
SD	3.87	0.51	0.49	0.49

Note: Male (male = 1, female = 0), Bryan (=1, others = 0) and Marijke (=1, others = 0) are dummy variables; n = 18.
*p < .05.

Table 5
Regression of Math Scores on Student Sex and Teachers

Variable	b	se	ΔR^2	95% CI	F	t
Male	5.00	0.49	.44	3.95, 6.05		10.25*
Teacher			.50		59.27*	
Bryan	3.00	0.60		1.72, 4.28		5.02*
Marijke	6.50	0.60		5.22, 7.78		10.88*
Intercept	72.50	0.49		71.45, 73.55		148.58*

Note: R^2 = .94, adj. R^2 = .93, $F_{3,14}$ = 74.51*, MSE = 1.071, n = 18. ΔR^2 represents the semi-partial correlation or the increment in R^2 due to adding the respective variable. Male (male = 1, female = 0), Bryan (=1, others = 0) and Marijke (=1, others = 0) are dummy variables.
*p < .05.

Table 6
Comparisons of Adjusted Mean Math Scores Among Teachers

Comparison	Estimated Mean Difference	Standard Error of Difference	Bonferroni Adjusted 95% CI
Bryan vs. Gunther	3.00*	0.598	1.38, 4.62
Marijke vs. Gunther	6.50*	0.598	4.88, 8.12
Bryan vs. Marijke	-3.50*	0.598	-5.12, -1.88

*p<.05, where p-values are adjusted using the Bonferroni method.

Regression results show that both student sex and teachers are statistically related to students' math scores at the .05 level of significance. Males score about 5 points higher than females, and students in Marijke's class tend to score higher than students in either of Bryan's or Gunther's class. Students in Gunther's class score lower than in either Bryan's or Marijke's class. Note that all teacher comparisons are statistically different.

8. Exercises

(1) According to the leadership literature, there are a number of different leadership styles. Listed below are scores obtained from an instrument designed to measure a particular leadership style, which will be referred to as style X. Of interest is whether X differs by school district type in terms of urbanity, and by sex. A stratified random sample of school principals were selected from three district types (mostly urban, mostly suburban, and mostly rural).

The scores on style X range from 100 to 0. The closely the score to 100, the more the respondent conforms to style X, while the closer the score to 0, the less the respondent conforms to style X.

Is there any evidence that X differs among the three district types, or by sex?

Sex	District Type	Style X
m	urban	85
m	urban	98
m	urban	75
f	urban	63
m	urban	91
f	urban	49
f	urban	62
f	suburban	49
f	suburban	48
m	suburban	56
m	suburban	78
f	suburban	35
m	suburban	50
m	rural	33
m	rural	95
f	rural	26
f	rural	11
f	rural	33
m	rural	25
m	rural	65

(2) A researcher is interested in learning whether frequency of reading at home to elementary-aged children produces differential effects on reading achievement. After obtaining information from a randomly selected sample of parents about this behavior, the following classifications and standardized achievement scores were recorded. (Note: frequency classifications as follows: a = less than once per month, b = once to three times per month, c = more than three times per month.) In addition to reading frequency, information regarding the family's status concerning whether or not the family's child receives either free or reduced lunch is recorded as a proxy for SES.

SES	Freq. of Reading	Achievement
fr	a	48
fr	a	37
no	a	47
no	a	65
no	b	57
fr	b	39
fr	b	49
no	b	45
no	c	61
no	c	55
fr	c	51
fr	c	30

Note. FR indicates free or reduced lunch received, NO indicates otherwise.

Is frequency of reading at home related to student reading achievement once SES is taken into account?

(3) An administrator wishes to know whether student behavioral problems can be linked to student performance. If students were suspended or reprimanded more than once, they are classified as having behavioral problems. In addition, each student's SES is known, and should be taken into account. The administrator randomly selects 13 students and collects the appropriate data.

Student	GPA	Student SES	Behavioral Problems
Bill	3.33	h	n
Bob	1.79	l	y
Stewart	2.21	m	n
Linda	3.54	h	y
Lisa	2.89	m	n
Ann	2.54	m	n
Fred	2.66	h	y
Carter	1.10	l	y
Bill	3.10	h	n
Sue	2.10	l	y
Kara	2.07	l	y
Loser	2.31	m	n
Kathy	3.67	h	n

9. Exercise Answers

(1) Results for leadership style analysis.

Table 1a
Descriptive Statistics for Leadership Style, District Type, and Sex

Variable	Correlations			
	Style	Urban	Suburban	Male
Style	---			
Urban	.55*	---		
Suburban	-.10	-.48*	---	
Male	.54*	.03	-.07	---
Mean	56.35	.350	.300	.550
SD	25.07	.489	.470	.510

Note: Male is a dummy variable (male = 1, female = 0), as are Urban (1, 0 = other) and Suburban (1, 0 = other); n = 20.

Table 1b
Regression of Style on Sex and District Type

Variable	b	se	ΔR^2	95%CI	F	t
Male	26.29	7.53	.29	10.32, 42.26		3.49*
District Type			.33		7.14*	
Urban	33.57*	8.94		14.62, 52.52		3.76*
Suburban	13.40	9.32		-6.36, 33.16		1.44
Intercept	26.12	7.65		9.91, 42.33		3.42*

Note: R^2 = .625, adj. R^2 = .555, $F_{3,16}$ = 8.90, MSE = 279.70, n = 20. ΔR^2 represents the semi-partial correlation or the increment in R^2 due to adding the respective variable. Male is a dummy variable (male = 1, female = 0), as are Urban (1, 0 = other) and Suburban (1, 0 = other).
*p < .05.

Table 1c
Comparisons of Style Scores Among Urban, Suburban, and Rural Principals

Contrast	Estimated Mean Difference	Standard Error of Difference	Bonferroni Adjusted 95% CI
Urban vs. Rural	33.57*	8.94	9.74, 57.40
Suburban vs. Rural	13.40	9.32	-11.44, 38.24
Urban vs. Suburban	20.17	9.32	-4.67, 45.01

*p<.05, where p-values are adjusted using the Bonferroni method.

[Note, Bonferroni CI taken from Excel Spreadsheet is incorrect so must calculate CI using tabled values for Bonferroni comparisons. Use male = .55 in regression equation to obtain estimated means for each district.]

Both sex and district type are statistically related to leadership style. Once district type is taken into account, males average about 26 points higher than females. Among the three district types considered, principals in urban settings have a statistically higher score on style than do principals in rural districts, but not statistically higher than principals in suburban districts.

(2) Results for reading frequency.

Table 2a
Descriptive Statistics for Achievement, SES, and Reading Frequency

Variable	Correlations			
	Achievement	B	C	SES
Achievement	---			
B = 1 to 3 per month	-.09	---		
C = more than 3 per month	.04	-.50	---	
SES	-.65*	.00	.00	---
Mean	48.66	.333	.333	.500
SD	10.129	.492	.492	.522

Note: SES is a dummy variable (free/reduced lunch = 1, otherwise = 0), as are B (1, 0 = other) and C (1, 0 = other); n = 12.

Table 2b
Regression of Achievement on Reading Frequency and SES

Variable	b	se	ΔR^2	95%CI	F	t
SES	-12.66	5.16	.426	-24.57, -0.77		-2.45*
Reading Freq.			.007		0.05	
B	-1.75	6.32		-16.32, 12.82		-0.28
C	-0.00	6.32		-14.57, 14.57		0.00
Intercept	55.58	5.16		43.68, 67.48		10.77*

Note: R^2 = .43, adj. R^2 = .22, $F_{3,8}$ = 2.04, MSE = 79.89, n = 12. ΔR^2 represents the semi-partial correlation or the increment in R^2 due to adding the respective variable. SES is a dummy variable (free/reduced lunch = 1, otherwise = 0), as are B (1, 0 = other) and C (1, 0 = other).
*p < .05.

Table 2c
Comparisons of Achievement among Reading Frequency

Contrast	Estimated Mean Difference	Standard Error of Difference	.95CI
B vs. A	-1.75	6.32	-16.32, 12.82
C vs. A	-0.00	6.32	-14.57, 14.57
B vs. C	-1.75	6.32	-16.32, 12.82

*p < .05.

[Note – the above comparison represents the unadjusted comparisons (no Bonferroni corrections); these numbers obtained from regression output. Bonferroni adjusted comparisons reported below in 2d.]

Table 2d
Comparisons of Adjusted Mean Reading Achievement Scores

Comparison	Estimated Mean Difference	Standard Error of Difference	Bonferroni Adjusted 95% CI
A vs. c	0.00	6.32	-18.99, 18.99
B vs. c	-1.75	6.32	-20.74, 17.24
A vs b	1.75	6.32	-17.24, 20.74

*p<.05, where p-values are adjusted using the Bonferroni method.

Only SES was statistically related to achievement scores, with those receiving free for reduced lunch scoring about 12 to 13 points lower than those not receiving free/reduced lunch, on average. There were no statistical differences observed among the three levels of reading frequency.

Bonferroni and Scheffe adjusted confidence intervals are reported below.

```
. regr achievement i.read_freq_num i.ses_num
. pwcompare read_freq_num, bonf

Pairwise comparisons of marginal linear predictions

----------------------------------------------------------------
              |                                Bonferroni
              |   Contrast   Std. Err.     [95% Conf. Interval]
--------------+-------------------------------------------------
read_freq_num |
      2 vs 1  |      -1.75   6.320436     -20.81093    17.31093
      3 vs 1  |   7.07e-15   6.320436     -19.06093    19.06093
      3 vs 2  |       1.75   6.320436     -17.31093    20.81093
----------------------------------------------------------------

. pwcompare read_freq_num, sch
----------------------------------------------------------------
              |                                 Scheffe
              |   Contrast   Std. Err.     [95% Conf. Interval]
--------------+-------------------------------------------------
read_freq_num |
      2 vs 1  |      -1.75   6.320436     -20.62467    17.12467
      3 vs 1  |   7.07e-15   6.320436     -18.87467    18.87467
      3 vs 2  |       1.75   6.320436     -17.12467    20.62467
----------------------------------------------------------------
```

(3) Results for GPA analysis.

Table 3a
Descriptive Statistics for GPA, SES, and Behavioral Problems

Variable	Correlations			
	GPA	High	Mid	Behavior
GPA	---			
High SES	.78*	---		
Mid. SES	-.07	-.53	---	
Behavior	-.46	-.10	-.62*	---
Mean	2.56	0.39	0.31	0.46
SD	0.74	0.51	0.48	0.52

Note: High (1, 0 = otherwise) and Mid. SES (1, 0 = otherwise) are dummy variables, as is behavior (1 for problems, 0 = otherwise); n = 13.

Table 3b
Regression of GPA on Behavioral Problems and SES

Variable	b	se	ΔR^2	95%CI	F	t
Behavioral Prob.	-.27	.37	.01	-1.10, 0.57		-0.72
SES			.56		11.31*	
High	1.34*	.35		0.54, 2.13		3.81*
Mid	.46	.47		-0.60, 1.51		0.98
Intercept	2.03	.42		1.08, 2.98		4.83*

Note: R^2 = .78, adj. R^2 = .70, $F_{3,9}$ = 10.35*, MSE = 0.164, n = 13. ΔR^2 represents the semi-partial correlation or the increment in R^2 due to adding the respective variable.
*p < .05.

Table 3c
Comparisons of Achievement among Reading Frequency

Contrast	Estimated Mean Difference	Standard Error of Difference	95% CI of Mean Difference
High vs. Low	1.34*	0.35	0.54, 2.21
Mid vs. Low	0.46	0.47	-.60, 1.51
High vs. Mid	0.88	0.31	.18, 1.58

*p < .05.

Only SES was statistically related to GPA, with those in the high SES group showing statistically higher GPAs than either the middle or low SES groups. There was no statistical difference between the middle and low SES groups, nor was behavioral problem associated with GPA.

[Table 3c above are the unadjusted comparisons, Table 3d below contains the Bonferroni adjusted comparisons using the estimated means with behavioral problems mean used as 0.46 to obtained predicted means for each of the three SES groups.]

[Again note that the Excel spreadsheet se are too small and erroneous, so use tabled Bonferroni critical t and calculate CI using regression se.]

Table 3c
Comparisons of Achievement among Reading Frequency

Contrast	Estimated Mean Difference	Standard Error of Difference	Bonferroni Adjusted 95% CI
High vs. Low	1.34*	0.35	0.31, 2.36
Mid vs. Low	0.46	0.47	-0.91, 1.83
High vs. Mid	0.88	0.31	-0.03, 1.79

*p<.05, where p-values are adjusted using the Bonferroni method.

[Bonferroni critical t = 2.923 (3 comparions, 9 df)]

Scheffé confidence intervals are reported below.

Table 3c
Comparisons of Achievement among Reading Frequency

Contrast	Estimated Mean Difference	Standard Error of Difference	Bonferroni Adjusted 95% CI
High vs. Low	1.34*	0.35	0.31, 2.36
Mid vs. Low	0.46	0.47	-0.91, 1.82
High vs. Mid	0.88	0.31	-0.02, 1.78

*p<.05, where p-values are adjusted using the Scheffé method.